École Normale Supérieure
FIMFA — Statistique 2017–2018

Devoir 4 pour le 8 janvier 2018

1. Théorie de Vapnik-Chervonenkis

Dans ce problème, on cherche à étendre les résultats obtenus sur la statistique de Kolmogorov-Smirnov à des classes d'événements plus générales. On se donne un sous-ensemble $\mathcal{C}$ de la tribu $\mathcal{F}$ définie sur Ω. On suppose que $(\Omega, \mathcal{F})$ est muni d'une loi P. Les variables aléatoires $X_1, X_2, \ldots, X_n, \ldots$ sont i.i.d. selon P. On note P_n la loi empirique : $\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$. On s'intéresse à

$$Z = \sqrt{n} \sup_{A \in \mathcal{C}} |P_n(A) - P(A)|.$$

Dans le cas de la statistique de Kolmogorov-Smirnov, $\mathcal{C}$ est constitué par l'ensemble des demi-droites $(-\infty, x]$ pour $x \in \mathbb{R}$. On

On appelle trace de $\mathcal{C}$ dans la suite $X_1^n = (X_1, \ldots, X_n)$ la collection de parties de $\{X_1, \ldots, X_n\}$ obtenues en intersectant $\{X_1, \ldots, X_n\}$ avec un élément of $\mathcal{C}$. Le coefficient d'éclatement de Vapnik-Chervonenkis VC $\mathbb{S}_{\mathcal{C}}(X_1^n)$ est le cardinal de la trace de $\mathcal{C}$ dans X_1^n.

(a) Montrer que pour toute fonction convexe croissante Ψ,

$$\mathbb{E}\left[\Psi\left(Z\right)\right] \leq \mathbb{E}\left[\Psi\left(\frac{2}{\sqrt{n}} \sup_{A \in \mathcal{C}} \left|\sum_{i=1}^{n} \epsilon_i \mathbb{I}_A(X_i)\right|\right)\right]$$

où $\epsilon_1, \ldots, \epsilon_n$ sont des variables de Rademacher indépendantes et indépendantes de $X_1, \ldots, X_n$.
Suggestion : consulter Cours Section 8.5

(b) Avec les mêmes notations que pour la question précédente, montrer que

$$\mathbb{E}\left[Z\right] \leq \kappa \times \mathbb{E}\left[\sqrt{\log\left(\mathbb{S}_{\mathcal{C}}(X_1^n)\right)}\right]$$

pour une constante universelle κ (qui ne dépend ni de n ni de $\mathcal{C}$).
Suggestions : consulter Cours Section 1.4, inégalité de Hoeffding et TD 2 Exercice 3.

(c) Toujours avec les mêmes notations, montrer que

$$\mathbb{E}\left[\Psi\left(Z\right)\right] \leq \mathbb{E}\left[\mathbb{S}_{\mathcal{C}}(X_1^n)\right] \times \mathbb{E}\left[\Psi\left(\left|\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i\right|\right)\right],$$

et

$$\mathbb{P}\left\{Z > x\right\} \leq 2\mathbb{E}\left[\mathbb{S}_{\mathcal{C}}(X_1^n)\right] \times \mathrm{e}^{-\frac{x^2}{8}}.$$

(d) Les deux bornes de la question précédente peuvent elles être améliorées lorsque $\mathcal{C}$ est formée par les demi-droites ?

(e) On a toujours $\mathbb{S}_{\mathcal{C}}(X_1^n) \leq 2^n$. On dit que $\mathcal{C}$ est une classe de Vapnik de VC dimension $v \in \mathbb{N}$ si et seulement si

$$v = \max\left\{n : \exists X_1^n \in \Omega^n, \mathbb{S}_{\mathcal{C}}(X_1^n) = 2^n\right\}.$$

Montrer que si $\mathcal{C}$ est une classe de Vapnik de VC-dimension v

$$\mathbb{S}_{\mathcal{C}}(X_1^n) \leq \sum_{k=0}^{v} \binom{n}{k}.$$

Vérifier que

$$\sum_{k=0}^{v} \binom{n}{k} \leq \left(\frac{en}{v}\right)^v .$$

(f) Considérons le cas où $\mathcal{C}$ est formé par les demi-espaces de $\mathbb{R}^d$: $\Omega = \mathbb{R}^d$, et $A \in \mathcal{C}$ si et seulement si il existe $w \in \mathbb{R}^d, b \in \mathbb{R}$ tels que

$$A = \left\{x : x \in \mathbb{R}^d, \langle w, x\rangle > b\right\} .$$

Montrer que la dimension de Vapnik-Chervonenkis de $\mathcal{C}$ est égale à $d+1$.

(g) Montrer que l'ensemble des convexes de $\mathbb{R}^2$ n'est pas une classe de Vapnik(-Chervonenkis).

2. Application de la théorie de Vapnik-Chervonenkis à la théorie de l'apprentissage

En classification binaire, les données proviennnent de tirages i.i.d. selon une loi inconnue P sur $\Omega \times \{0,1\}$, P est la loi jointe de (X,Y) avec X à valeur dans Ω et Y à valeur dans $\{0,1\}$. Un classifieur g est une fonction (mesurable pour la tribu sous-entendue sur Ω) de Ω dans $\{0,1\}$. Le risque en prédiction du classifieur g, noté $R(g)$ est défini par

$$R(g) = P\{g(X) \neq Y\} = \mathbb{E}_P\left[\mathbb{I}_{g(X)\neq Y}\right] .$$

En apprentissage on cherche à construire à partir d'un échantillon D_n i.i.d. selon P ($D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$), un classifieur g_n dont le risque $R(g_n)$ soit aussi petit que possible. La difficulté vient de ce que P n'est pas connue.

(a) Montrer qu'il existe un classifieur g^* qui minimise $R(g)$. Caractériser ce classifieur à partir de $\mathbb{E}[Y \mid X]$. Pourquoi appelle-t-on ce classifieur le classifieur de Bayes ?

Pour « apprendre » un classifieur $\widehat{g}_n$, on se donne un dictionnaire $\mathcal{H}$ de fonctions de $\Omega \to \{0,1\}$ et on cherche à minimiser le risque empirique

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{g(X_i)\neq Y_i} = P_n\{g(X) \neq Y\} .$$

sur $\mathcal{H}$:

$$\widehat{g}_n = \mathrm{argmax}_{g\in\mathcal{H}}\, R_n(g) .$$

(b) On suppose que $\mathcal{H}$ définit une classe de Vapnik-Chervonenkis de dimension v. Donner un majorant de

$$\mathbb{E}\left[R(\widehat{g}_n) - R_n(\widehat{g}_n)\right]$$

et un majorant (non trivial) de

$$\mathbb{P}\left\{R(\widehat{g}_n) - R_n(\widehat{g}_n) \geq x\right\} .$$

(c) Application aux classifieurs linéaires. Pour $\Omega = \mathbb{R}^d$, $\mathcal{H}$ formés par les fonctions de la forme $g_{w,b}(x) = \mathbb{I}_{\langle w,x\rangle > b}$, $w \in \mathbb{R}^d, b \in \mathbb{R}$.